SM McGhee 麥潔儀
J Brazier
CLK Lam 林露娟
LC Wong 黃麗君
J Chau 周宗欣
A Cheung 張惠棱
A Ho 何麗儀

Quality-adjusted life years: population-specific measurement of the quality component

Key Message

A valid scoring algorithm was developed to translate local SF-36 datasets to quality-adjusted life years.

Introduction

The ever-increasing demand on health care services results in a growing demand for appropriate methods of measuring and valuing the benefits of health care interventions (cost-effectiveness) in order to formulate policies. In Hong Kong, there is little explicit information to guide policymaking on which treatments to offer and what priority to allocate to differing sectors of the health care system.

A principal approach for incorporating preferences into a measure of health has been to value health status in a single unit of measurement known as quality-adjusted life years (QALYs),[1] which combines increased life expectancy and improvements in health status. This assigns to each period of time a 'quality weight' ranging from 0 to 1, which corresponds to the health-related quality during that period, where a weight of 1 is given to optimal health, and 0 to a health state equivalent to death. The number of QALYs is the value given to each state multiplied by the length of time spent in that state; thus a person expected to survive 10 years at a mean quality of 0.8 has 8 QALYs.

The Short Form-36 (SF-36) is a measure of perceived health status evolved out of two major research programmes in the USA.[2,3] It has become one of the most widely used measures of health status in clinical trials throughout the world. The SF-36, and its shorter version, the SF-12, have been assessed for relevance, translated and validated in Hong Kong.[4-9] This local version (SF-36 HK) is also widely used. In the United Kingdom, valuation data were collected, and a model was estimated to allow the calculation of a preference-based index for the English version of the SF-36 in its population.[10] Nonetheless, there is a major scientific concern when using UK preference weights in other countries.

We aimed to derive an algorithm to translate SF-36 data to utility weights for use in Hong Kong. The objectives were (1) to use a representative sample of the local population to obtain a series of valuations of health states that are locally relevant, based on the Hong Kong Chinese version of the SF-36 (SF-36 HK); (2) to use these valuations to derive a model which can be used to predict the value of any health state described by the SF-36 HK; and (3) to compare the final model results with the model already derived for the UK population to identify any systematic differences in valuations.

Methods

This study was conducted from 30 September 2004 to 30 June 2006. We valued a small number of health states, which could be extrapolated to all health states described by the SF-36. These health states were described by the SF-6D, a subset of the SF-36. The first step in creating the Hong Kong-based utility weights was to develop the SF-6D HK.

The SF-6D asks about six aspects of health (physical functioning, role limitations, social functioning, pain, mental health, and vitality) over the past 4 weeks. A health state is defined by taking one level from each of the six different aspects of the SF-6D, so each health state is described by six digits. The Chinese Hong Kong version of the SF-6D was derived from the English UK version by

Hong Kong Med J 2011;17(Suppl 6):S17-21

The University of Hong Kong: School of Public Health
SM McGhee, LC Wong, J Chau, A Cheung, A Ho
Department of Family Medicine and Primary Care
CLK Lam
University of Sheffield, United Kingdom
J Brazier

HHSRF project number: 02030351

Principal applicant and corresponding author:
Prof SM McGhee
School of Public Health, The University of Hong Kong, Pokfulam, Hong Kong SAR, China
Tel: (852) 2819 9280
Fax: (852) 2855 9528
Email: smmcghee@hku.hk

forward-backward translations. The SF-6D HK was field-tested in a pilot study and showed that the valuation method was feasible and the resulting data were reliable and fitted quite well in an econometric model.[11]

A representative sample was obtained through a random digit telephone survey between 17 October 2004 and 23 December 2005. Cantonese speaking residents aged ≥18 years were included. When a household was identified, a Kish Grid method was used to target a random respondent from a list of household members ranked by age. That person was asked for, and if unavailable an arrangement was made to call him/her back at a more convenient time. When the target respondent was contacted, some initial data (SF-6D, age, sex, educational level, living district, smoking status) were collected. An arrangement was made for a face-to-face interview to carry out the standard gamble procedure. To encourage participation, the face-to-face interviews were held in local community halls at convenient times of the day and evening, including weekends. A letter stating the date, time, and location of the interview was sent to each respondent 2 weeks earlier, and a reminder call was made the night before. Any respondent who did not show up was called for. Those who participated were given HK$100 for travelling expenses.

The composition of the final sample of face-to-face interviews was monitored with regard to sex, age, and district of residence. In the later stage of recruitment, there were insufficient respondents in the youngest age group (18 to 39 years). Therefore, recruitment of this group continued while recruitment of the other age groups had ceased.

Selection and valuation of health states

There were 196 health states selected as being able to be extrapolated to the full set of 18 000 states described by the SF-36. Each respondent was asked to value seven health states in a random order. In order to ensure each person valued a range of health states from very mild to very severe, the states were stratified into a block system.

The interview procedure was modelled on that used in the UK study, which was based on methodology developed at McMaster University, Canada. Each participant was asked to rank a set of ten health state cards and then rate them using a visual analogue scale of 0 to 100 points, where the endpoints were the best and the worst imaginable health states. This ranking and rating exercise was followed by the standard gamble procedure. While evaluating each of the seven health states, the participants were asked to choose between an intervention (choice A) involving uncertain health outcomes and a certain health state defined by the SF-6D HK states (choice B). There were two possible health outcomes in choice A: the best health state (H) if the treatment was successful and the worst health state (K) if the treatment failed. The seven health states were placed in choice B one by one as the certain health state under valuation. Once the seven states had been valued, an eighth choice was presented. The reference state in choice A was represented by the state judged by this individual as the poorest, either K or L. The other one of states K or L then became choice B. This enabled valuation of K against L.

Choice A, the intervention, involved the best health outcome with probability P and the worst health outcome with probability $1–P$. The probability of the outcomes was varied until the respondent was indifferent between choosing A or B. At that point, the preference-based utility value of the health state in choice B was reached. The respondent was asked for further information on marital status, number of children, employment, place of birth, number of family members, type of housing, household income, medical benefits, and health-related questions in order to control for these in the analysis. The value for each health state was then transformed to a scale with 1 as full health and 0 equivalent to death.

The modelling methods used were the same as for the UK study.[12] Two main modelling approaches were used with either individual level data, which takes into account the variation across respondents using a random effects model or mean level data, in which explanatory variables were used to estimate the mean value given to each state. A set of binary dummy variables was created to describe each level and dimension of the health state. There was a binary dummy variable to take account of any additional effect on the health state value when one or more dimensions of health were at the most severe level. The models were estimated by the ordinary least square mean level model with constant forced through unity and by maximum likelihood for the random effects model, with the most severe term included to account for interactions. Explanatory power for the ordinary least square model was expressed in terms of an adjusted R-squared. The ability of the models to predict health state values was assessed in terms of mean absolute errors and the proportion of predictions outside 0.05 and 0.10 on either side of the actual value. Predictions were further tested in terms of bias using t-tests. Since the levels of each dimension were ranked progressively worse, the dummies represented progressively worse problems compared to a baseline with no problem for that dimension. Therefore, the coefficient estimates for the dummies on each dimension should be negative and increasing in size. An inconsistent result was one where a coefficient decreased rather than increased in size. All analyses were carried out in STATA 8.2.

Results

Over 14 months, 16 400 telephone calls were made, from which 6746 potential households were identified. Of these, the targeted individual was working overseas, had hearing or language problems precluding a telephone interview, could not be contacted after 10 attempts or another household member refused on their behalf, leaving 2544 targeted respondents who were contacted. Of these, 392 either

refused or gave incomplete responses and 2152 completed the initial telephone survey, giving a response rate for the telephone survey of 85% (2152/2544) of contacted target respondents or 32% (2152/6746) of possible cases.

Of the 2152 respondents who completed the telephone interview, 964 (45%) agreed to participate in the face-to-face interviews, of which 641 (66%) were eligible, willing and able to complete the interviews. Sex, age, living district, and smoking habit were similar between participants and non-participants, but the latter had lower educational qualifications (Table 1).

More severe role limitation was reported by participants and poorer social functioning by non-participants. Of the 641 participants in the face-to-face interviews, 29 (4.5%) were excluded because they failed to value the pits state and were therefore unable to generate an adjusted standard gamble value. A further 30 participants who gave the same valuation for each of the seven intermediate health states were also excluded, leaving 582 participants' data for analysis; each made eight standard gamble valuations giving 4656 valuations, of which 60 (1.3%) were illogical so 4596 valuations were finally analysed.

Table 1. Characteristics of participants and non-participants in the face-to-face interview

Variable	No. (%) of non-participants (n=1496)	No. (%) of participants (n=575)	P value, Chi-square test
Male	639 (43)	227 (40)	0.178
Age (years)			0.112
18-39	732 (49)	291 (51)	
40-64	486 (33)	201 (35)	
≥65	269 (18)	82 (14)	
Education level			<0.001
Tertiary	372 (25)	203 (36)	
Secondary	711 (48)	270 (47)	
Primary	276 (19)	84 (15)	
None	122 (8)	15 (3)	
Living district			0.078
Hong Kong	202 (14)	102 (18)	
Kowloon	481 (33)	176 (31)	
New Territories	775 (53)	296 (52)	
Smoking habit			0.177
Never	611 (78)	251 (83)	
Current	123 (16)	36 (12)	
Ex-smoker	49 (6)	15 (5)	

Table 2. Sample representativeness

Variable	% of sample (n=582)	% of population	Effect size*
Sex			0.19
Male	37.8	47.3	
Age (years)			0.18
18/15-24†	21.1	15.3	
25-34	15.8	17.7	
35-44	22.3	22.3	
45-54	16.2	19.9	
55-64	11.7	10.7	
≥65	12.9	14.1	
Highest education level			0.46
Primary	13.7	25.9	
Secondary	48.3	51.7	
Tertiary (non-degree)	17.4	07.6	
Tertiary (degree)	20.6	14.8	

* Effect size of 0.1=small, 0.3=medium, and 0.5=large
† Data for population are ≥15 years and for sample are ≥18 years

Comparisons were made between the final sample of respondents and the 2005 population[13] on sex, age and education (Table 2). The effect sizes for the variables of sex and age were 0.19 and 0.18, respectively, which were small and the sample was reasonably representative in these variables. However, the sample included more highly educated respondents than the general population, with a medium-to-large effect size of 0.46. Thus, the impact of weighting the sample by education level was examined.

Health state values

There were 158/197 (80%) health states with a median value greater than the mean, indicating that the data were skewed left.

Individual model

All beta coefficients had the expected negative sign in the model and were significant at the 10% level. There were three inconsistent coefficients: pain 2 to 3, pain 2 to 4, and mental health 2 to 3. In each case the higher level should have had a larger negative coefficient but did not. The UK model had four such inconsistencies. There was also evidence of some bias ($t \neq 0$) in the predictions of the random effects model, as there was in the UK study. However, the overall predictive ability was good with a mean absolute error of 0.070 compared to 0.078 in the UK study.

Mean model

All coefficients were significant and there were four inconsistencies: physical functioning 2 and 3, social functioning 2 and 3, pain 2 and 3, and mental health 4 and 5. For two of these, the difference was very small (0.001) and none were greater than 0.02. The predictive ability of the mean model was better than the individual model with a mean absolute error of 0.057 (compared to 0.075 in the UK study). As the mean was an ordinary least squares model, it was unbiased.

The performance of the Hong Kong models compared favourably with the UK models in terms of the mean absolute error and the number of absolute errors greater than 0.05 or 0.10. The results supported the validity of preference-based valuation of the SF-6D HK in the local population.

Improving the Hong Kong model

In order to generalise our model, weights were created by dividing the population proportion by the sample proportions by age, sex, and education. These weights were incorporated into the model in STATA. The coefficients from the weighted and unweighted models were similar, as were the mean absolute error and the number of absolute errors greater than 0.05 or 0.10.

Table 3. Coefficients for the consistent version of the mean model

SF-6D item*	Coefficients for the mean consistent model†
PF2	-0.050
PF3	-0.056
PF4	-0.092
PF5	-0.103
PF6	-0.178
RL2	-0.035
RL3	-0.035
RL4	-0.054
SF2	-0.039
SF3	-0.050
SF4	-0.050
SF5	-0.073
PAIN2	-0.037
PAIN3	-0.037
PAIN4	-0.052
PAIN5	-0.060
PAIN6	-0.100
MH2	-0.038
MH3	-0.058
MH4	-0.088
MH5	-0.088
VIT2	-0.039
VIT3	-0.056
VIT4	-0.063
VIT5	-0.077
Most severe level	-0.115

* PF denotes physical functioning, RL role limitations, SF social functioning, MH mental health, and VIT vitality

† All are significant at $t_{0.10}$

Final model for Hong Kong data

For the purpose of generating a model for use in cost utility analyses, the intercept to unity was restricted. The mean models appeared to be better than the individual models. The mean level model with the interaction term was therefore recommended. To deal with the few inconsistent values, the model was re-run to produce the consistent model (Table 3). To estimate the utility of a health state, these results were combined with a set of SF-36 data, and the specific health state can be described in terms of scores on the dimensions (Table 3). For example, if physical functioning scores level 2 (ie PF2) but all other dimensions score 1 (ie the best score), the utility is estimated by subtracting the coefficient 0.05 from 1, giving 0.95. If other dimensions score other than 1, then the relevant coefficients are combined to estimate the utility value of the health state.

Discussion

This study aimed to derive utility weights using a local population sample in such a way that they can be used to predict the value of any health state described by the locally validated SF-36. A standard gamble method was used to obtain the utility values in interviews with a final sample of 582 people. The sample underrepresented persons with lower levels of education but weighting made no appreciable difference to the results. Each individual valued a range of states of different predicted values so that further biases are unlikely.

In the standard gamble exercise, individuals were allowed time to understand and feel comfortable with the tasks. The proportion of participants who considered the quality of their answer poor or very poor was <2%, as was the proportion rated by the interviewer as not having understood the task. In the final model, some levels were re-grouped to produce a consistent algorithm for use. These inconsistencies should be further investigated to determine whether there are any implications for the wording of the SF-36 HK.

The mean model was considered as having better predictive ability, as the mean absolute error and number of absolute errors >0.05 or >0.10 were fewer when compared to the model with individual level data. The interaction terms were significant in both models and improved the predictive ability. In the comparison between the Hong Kong and UK models, there appeared to be some systematic difference in valuation in the two populations, for which more research is needed. A few interactions were also identified, but further investigation was limited because the number of observations required to do this exceeded the data available.

The results of this study provide a way of estimating a preference-based single index of utility for the health state of a Hong Kong sample. It gives an alternative to single index measures like the Health Utilities Index for those who would prefer to use the SF-36. It can also be applied retrospectively to existing SF-36 datasets.

Conclusions

Using the derived coefficients for the SF-6D health states, we can transform any set of SF-36 HK data to utility weights for the determination of QALYs. These QALYs are locally relevant and have a reasonable degree of validity. This enables calculations of costs per QALY for procedures and cost-utility analyses in any studies by using the SF-36 HK measures.

Acknowledgements

This study was supported by the Health and Health Services Research Fund (#02030351), Food and Health Bureau, Hong Kong SAR Government. We thank Prof A O'Hagan and Dr Samer A Kharroubi of the Centre for Bayesian Statistics in Health Economics, Department of Probability and Statistics, University of Sheffield for assistance in the selection of the health states to be included in the valuation exercise. We also thank those who have participated in the study.

References

1. Drummond MF, Stoddart GL, Torrance GW. Methods for the economic evaluation of health care programmes. Oxford: Oxford University Press; 1987.
2. Brook RH, Ware JE Jr, Rogers WH, et al. Does free care improve

adults' health? Results from a randomized controlled trial. N Engl J Med 1983;309:1426-34.
3. Tarlov AR, Ware JE Jr, Greenfield S, Nelson EC, Perrin E, Zubkoff M. The Medical Outcomes Study. An application of methods for monitoring the results of medical care. JAMA 1989;262:925-30.
4. Lam CL. Reliablity and construct validity of the Chinese (Hong Kong) SF-36 for patients in primary care. Hong Kong Pract 2003;25:468-75.
5. Lam CL, Chan MS, Ren XS, Gandek B. Chinese (Hong Kong) translation of the MOS 36-item short form health survey (SF36) standard version 1. October 1996. Revised May 1998.
6. Lam CL, Lauder IJ, Lam DTP, Gandek B. Validation and norming of the MOS 36-item Short Form Health Survey in Hong Kong Chinese adults. Health Services Research Committee Dissemination Report No. 711026. Hong Kong: Health Services Research Committee; November 2000.
7. Lam CL, Fong DY, Lauder IJ, Lam TP. The effect of health-related quality of life (HRQOL) on health service utilisation of a Chinese population. Soc Sci Med 2002;55:1635-46.
8. Lam CL, Tse EY, Gandek B. Is the standard SF-12 Health survey valid and equivalent for a Chinese population? Qual Life Res 2005;14:539-47.
9. Lam CL, Tse EY, Gandek B, Fong DY. The SF-36 summary scales were valid, reliable, and equivalent in a Chinese population. J Clin Epidemiol 2005;58:815-22.
10. Brazier J, Usherwood T, Harper R, Thomas K. Deriving a preference-based single index from the UK SF-36 Health Survey. J Clin Epidemiol 1998;51:1115-28.
11. Lam CL, Brazier J, McGhee SM. Valuation of the SF-6D health states is feasible, acceptable, reliable, and valid in a Chinese population. Value Health 2008;11:295-303.
12. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ 2002;21:271-92.
13. Hong Kong Census and Statistics Department. Women and men in Hong Kong: key statistics, 2006.